Redesigning National School Surveys: Coverage and Stratification Improvements using Multiple Datasets

William Robb, Kate Flint, Ronaldo Iachan

ICF International Inc.

40 Wall Street, New York, NY/william.robb@icfi.com

Abstract

This paper discusses the redesign of two national school surveys with a focus on sampling frame development. Historically, the sampling frames for these surveys have been developed from files acquired from a commercial vendor. Vendor files are used as they contain up-to-date contact information, facilitating recruitment efforts in support of high response rates. The vendor provided data set incorporates several sets of variables sourced from the NCES Common Core Data (CCD) and Private School Survey (PSS).

This paper explores the construction of an augmented frame, build by combing the vendor list with NCES files directly, using schools listed on the CCD for public schools and the PSS for non-public schools. A particular challenge is assembling an unduplicated frame for non-public schools in the absence of a unique identifier common across datasets. In addition, with multiple sources for both districts and schools, identifying a unique set of school-district relationships is difficult.

We provide an assessment of eligibility rates from each data source, and assess coverage improvements by grade, school type and geography.

Study Description

Both the School Health Policies and Programs Survey (SHPPS) and the National Youth Tobacco Survey (NYTS) are school based surveys that employ similar sampling designs, ultimately based on a national sampling frame of schools.

The 2014 National Youth Tobacco Survey (NYTS) employs a repeat cross-sectional design to develop national estimates of tobacco use behaviors and exposure to pro- and anti-tobacco influences among students enrolled in grades 6-12. The study represents the continuation of the NYTS cycles that took place in 1999, 2000, 2002, 2004, 2006, 2009, 2011, 2012 and 2013 conducted by ICF.

The primary objectives of the NYTS are to develop estimates of tobacco use behaviors and exposure to pro- and anti-tobacco influences among students enrolled in middle school and high school grades; to identify differences related to demographic characteristics (age, grade, gender, and race/ethnicity); and to determine whether there are observed trends over time in tobacco use behaviors and exposure to influences that promote or discourage tobacco use.

SHPPS is a study of school health-related policies at state, school district and school levels in K–12, conducted by CDC every six years since 1994 (1994, 2000, 2006 and 2012). The 2006 SHPPS assessed eight components of school health programs: school policy and environment, health education, physical education, health services, mental health and social services, food service, faculty, and staff health promotion, and family and community involvement.

Both studies employ multi-stage designs with stratification of both PSU, formed from districts for SHPPS, and Counties for NYTS, and schools. The final stage of sampling for SHPPS is schools and students for NYTS. The NYTS covers middle and high schools – or students in grades 6 – 12, and the SHPPS covers elementary, middle and high schools.

For the 2014 cycle of both SHPPS and NYTS, the basis of the sampling frame was expanded to improve frame coverage. Increases in coverage – that is, the proportion of the population under study that is represented in the sampling frame- reduce the risk of coverage bias. We used two lists of schools maintained by the National Center

for Education Statistics (NCES) to augment the list of schools, maintained by a private vendor, which we have used as the basis of the sampling frame since the inception of both NYTS and SHPPS.

While providing complete coverage is paramount, we also want the frame to exclude portions of the population that are ineligible. This means ensuring that all schools on the frame meet eligibility criteria in terms of grade ranges and other characteristics. Finally, each element of the frame should be represented only once. Meeting these criteria – removing duplicate records – was the central challenge in building the combined sampling frame from several source files.

A second concern was eligibility rates or over-coverage. Schools are screened for eligibility after sampling – and while the criteria are relatively broad, screening does represent a significant effort on the part of data collection staff. We wanted to be sure that the schools added to the frame did not decrease eligibility rates to such an extent that the screening process became unwieldy.

School Frame Construction

The vendor, or "MDR" file, utilized because of its up-to-date contact information, contains data fields from the NCES/CCD file described below to supply enrollments by grade and student distributions by race/ethnicity. These data fields were merged onto the MDR file and provided by the vendor, and support sample stratification, school classification, and computation of the measure of size used in the PPS sampling of schools employed by the NYTS.

For the prior cycles of both NYTS and SHPPS this file was the sole source of the sampling frame. In 2014, we explored a new frame development process that made use of a second, composite source for the sampling frame based on two data files available from the National Center for Education Statistics (NCES).

The first step in the frame build process was to pre-process each of the source files, constructing the necessary sampling and matching variables, and ensuring that coding of these variables was consistent across the files. Then, we combined files, removing duplicate records in the process, so that the resulting sampling frame has only one record per school.

There were a number of screening, or filtering, steps. First, the individual files were screened to ensure that they contained only schools in the 50 states and the District of Columbia, and that the schools did not serve a student population that was enrolled on a full time basis elsewhere.

After the files were combined, we filtered out ineligible schools; those that did not contain students in grades 6 – 12 for NYTS and those that did not contain students in grades 1 – 12 for SHPPS.

Two NCES datasets were used to assemble one frame representing the universe of public and private schools – the Private School Survey Frame and the Common Core of Data. These were then matched against the second frame representing this same universe – the vendor provide frame used as the sole frame in prior cycles of these studies.

The purpose of the Private School Survey (PSS) is to build an accurate and complete list of private schools to serve as a sampling frame for the NCES sample surveys of private schools, and to report the total number of private schools, teachers and students in the survey universe. It is conducted every two years.

The NCES defines a private school as one that is not supported primarily by public funds, provides classroom instruction for one or more of grades K-12 or comparable ungraded levels, and has one or more teachers.

Organizations or institutions that provide support for home schooling without offering classroom instruction are not included. The initially developed frame is updated by matching by lists provide by nationwide private school associations, state departments of education, and other national private school guides and sources. An additional frame search is conducted by the Bureau of the Census.

The NCES Common Core of Data (CCD) is a program that annually collects fiscal and non-fiscal data about all public schools, public school districts and state education agencies in the United States. Data are supplied by state education agency officials. The data is compiled from five surveys sent to state education departments, with most of the data obtained from administrative records, and covers approximately 100,000 public elementary and secondary

schools, 18,000 public school districts in the 50 states, District of Columbia, Department of Defense schools and outlying areas.

The vendor file is a list of all schools obtained from Market Data Retrieval, a leading U.S. provider of education marketing information services. MDR maintains a comprehensive database of educational institutions, updated on a rolling basis throughout the year. ICF obtains limited use-licenses for data records, including associated district and personal information, for all schools that contain any of grades 1 – 12.

We use an extract from their master database obtained in March of 2012.

All three files were pre-screened prior to linking to ensure comparable sets of schools. The MDR file was subset to include only building records. The CCD file was subset to include only open schools, excluding reportable programs and schools outside geographic scope. Finally the PSS file was screened to include only schools that had grades higher than Kindergarten.

Following linking, schools were screened for eligibility. Only schools containing students in grades 6 through 12 were include in the frame.

Combining Files

In order to eliminate duplicate listings for a school, records from each source file were linked, and only one entry was taken for each school. Due to the structure of the NCES files, the process was split between public and private schools, with the MDR file being divided into public and private schools. These MDR record sets were linked with the NCES/CCD and NCES/PSS files respectively.

For public schools, 90% of the records on the MDR file had a NCES ID. These schools were linked directly using this ID. We then attempted to link schools by telephone number, and finally by address. Schools that shared addresses or telephone numbers were excluded from this step, as they would not produce a unique linkage.

For private schools, there was no common ID. In addition, the public use PSS file did not contain telephone numbers. So, for private schools, we linked on address only.

Addresses were standardized using a mail preparation utility prior to matching.

Results

A detailed synopsis of record counts for each stage in the process performed on the NYTS frame is presented in Appendix A. The first section of the table gives record counts for each of the source files, both before and after pre-screening described above.

The second section of the table summarizes the linking process for public and private schools, accounting for records input and output to the linking process.

For public schools, 84% of the resulting 103,596 schools were common across both files, with the NCES/CCD file supplying 9,732 (10%) unique schools, and the MDR file supplying 6,260 (6%) unique schools. For private schools, 57% of the 29,303 schools were common across both files, with the NCES/PSS file supplying 8,094 (27%) unique schools and the MDR file supplying 5,389 (16%) unique schools (Figure 2).

Finally, after filtering the file for schools that contained students in grades 6 – 12, the frame contained 86,180 schools. 76.3% of these were common across both MDR and NCES sources, with the MDR file contributing 3,936 (9.1%) unique schools, and both NCES files contributing 12,595 (14.6%) unique schools.

Coverage Assessment

Coverage improvements are measured by the percentage increase in frame counts for both students and schools. For public schools, overall, coverage increased by 12%, and for non-public schools, by 39%. While coverage increases varied across school levels (high school, middle school) and across sampling strata, as detailed below, this pattern remained – the greatest gains were for non-public schools.

Increases in coverage were higher for high schools, at 15% and 46% for public and non-public schools respectively, than for middle schools, at 35% and 10%. At the student level, coverage increases were very close for both high schools and middle schools for both public schools, at 16%, and non-public schools, at 2%.

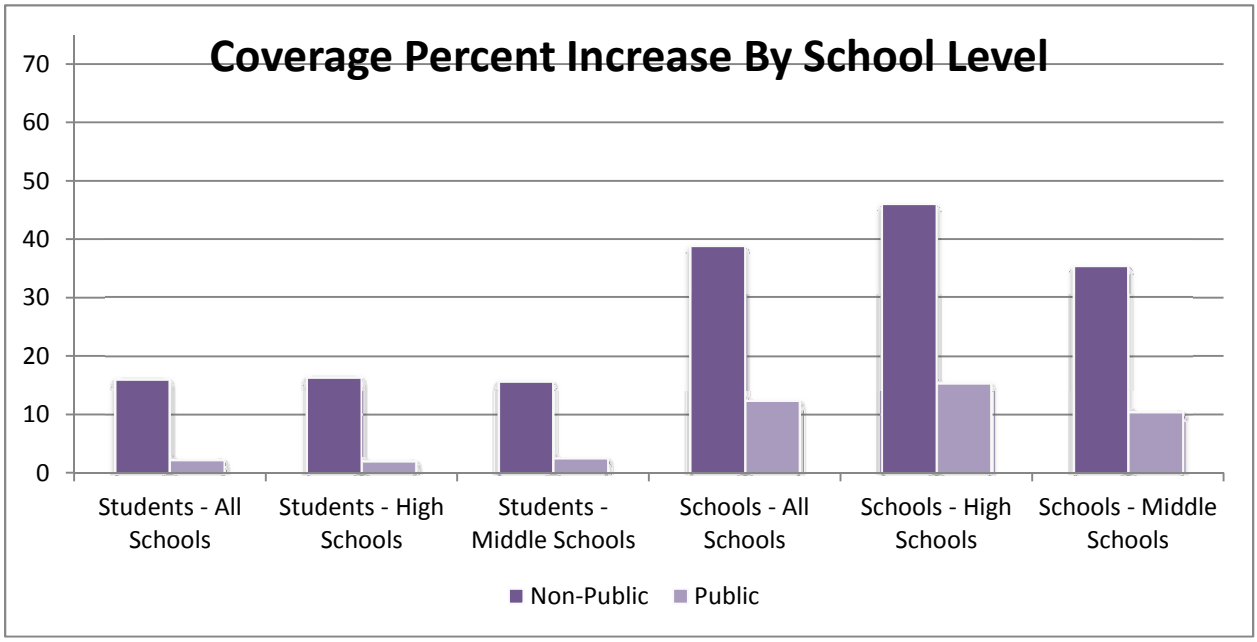
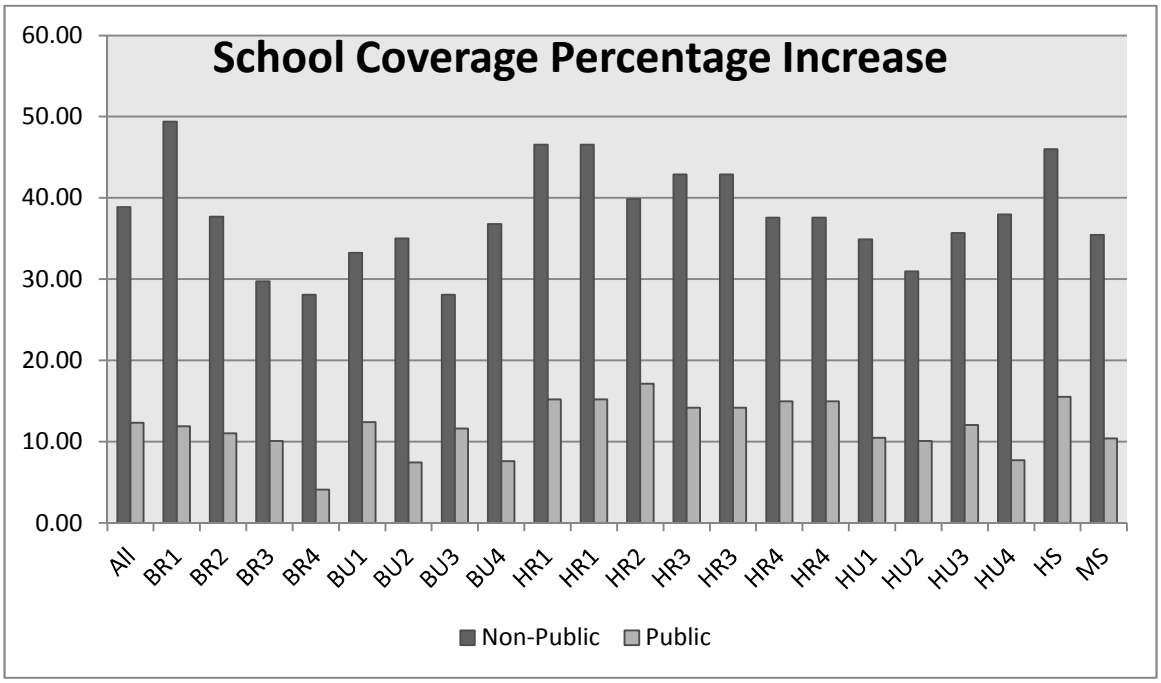


Figure 2 displays the coverage increases across sampling strata. Increases in coverage range from 49% (BR1) to 28% (BU3) for non-private schools, and from 17% (HR2) to 4% (BR4) for public schools.

When assessed at the student level – that is, looking at the number of students added, rather than the number of schools added, the increases in coverage are much lower, at 2% for public schools and 16% for private schools overall. Increases in coverage were much more similar across high- and middle-schools, differing by a percent or less. As with school coverage increases, there was considerable variability across sampling strata.

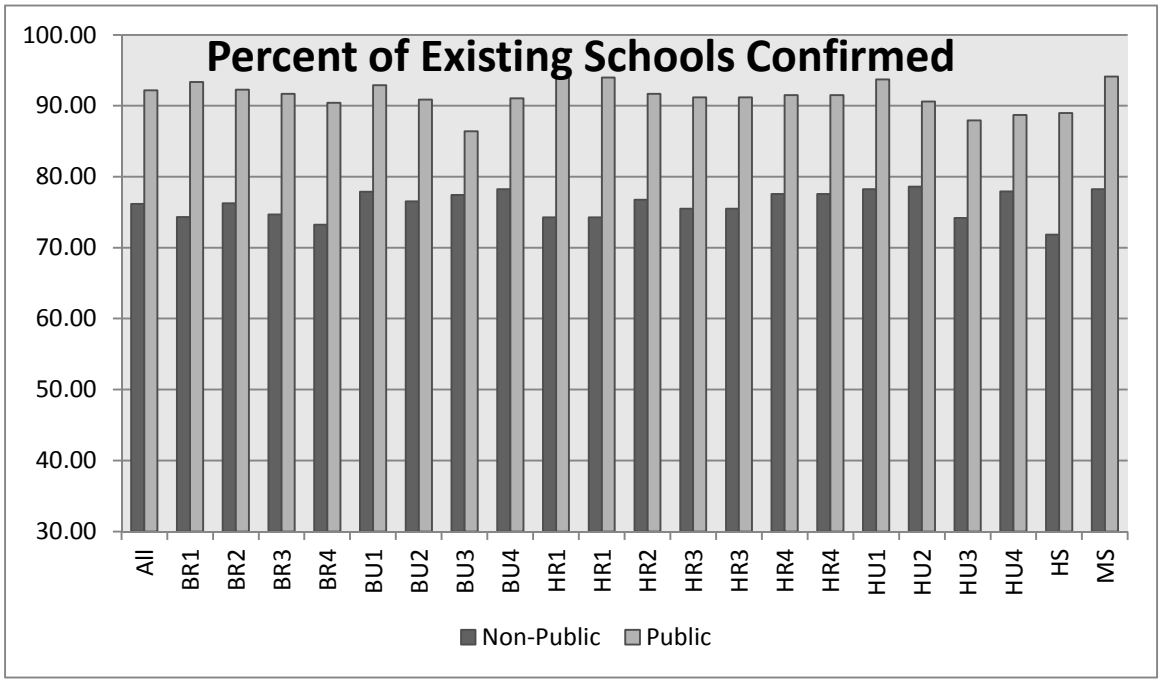


School Validation
Schools that were present on both files – that is, places were the data sources overlap – represent verification of the schools from two independent sources.

Confirmation rates were quite high for public schools, with 92% of school entries confirmed overall, 94% for middle schools, and 89% for high schools. Confirmation rates were not that variable across first stage strata, ranging from 86% to 94%. Rates were lower for private schools, with 76% of schools confirmed overall, and 72% and 78% of schools confirmed for high- and middle schools respectively. Validation rates ranged from 79% to 73% across first stage strata.

Duplication Assessment
It is possible that the linking procedures did not remove all duplicates from the combined frame. Assessing missed duplicates is a time consuming process, as it requires reviewing data records "by hand", examining addresses and school names for unmatched records, looking for names that differ slightly due to different usage, abbreviation or typographical errors.

Zip Codes were used as the unit of analysis for duplication assessment, with the number of added, overlapping (confirmed) and base (MDR) schools tallied by Zip Code. For Zip Codes with 100% overlap – that is, where all



schools were on both data sources, there was no possibility of duplicate schools, as all had been accounted for on both frames. For Zip Codes with less than 100% overlap, only areas with both new and base schools could contain overlaps. That is, Zip Codes that came into the frame as all schools in them were new, and Zip Codes where there was no added coverage have no potential for duplicate schools.

Considering only public schools less than 20% (table 1) showed a potential for un-linked duplicate schools. The vast majority of zip codes had no schools added, with a small percent in the opposite category – all schools in the Zip Code were new. Considering non-public schools, the percentage of Zip Codes that had potential duplicates was higher, at some 30%.

Table 1 - Potential for Duplicate Schools				
	Public Schools		Private Schools	
	Zip Code Count	Percent	Zip Code Count	Percent
Potential for Duplicates	3,822	18.9	3,131	29.9
No Duplicates - New Zip Code	16,018	79.3	5,993	57.1
No Duplicates - No Coverage Increase	349	1.7	1,366	13.0

Discussion & Conclusion

The overall goal of building a composite frame from multiple sources was to increase frame coverage. Based on frame processing counts we can see that the NCES data added some 12,500 schools, representing a 17% increase in schools listed on the frame. The increase was higher for private schools, with a 37% increase than for public, with a 10% increase.

Student coverage increases are substantially smaller than coverage increases at the school level. Coverage increases are primarily among smaller schools. For the NYTS, which employs no minimum school size criteria, this did not represent an increase in ineligible schools. However, it does have the potential to lower overall yields by increasing the likelihood that smaller schools will be included in the sample.

Finally we note that, despite the addition of a large number of schools, basic student populations with respect to sampling strata seem to be relatively stable. This indicates that the increased coverage was spread more or less uniformly across the frame.

While this certainly represents an increase in coverage, these results should be interpreted with caution. For one thing, we have no "gold standard" which would represent 100% non-duplicated coverage. If we take either complete set of schools – that represented by the MDR file or that represented by the combined NCES/PSS and NCES/CCD files as "truth", than we have overcoverage in the combined list. Specifically using the lists supplied by National Center for Education Statistics as the "gold standard", which is their stated purpose, we have 6% overcoverage for public schools and 19% for private schools.